CREATE COMPELLING GRAPHS AND FIGURES

35

Modern graphs are largely an eighteenth-century creation. Edward Tufte, the modern master of graphical display, states that Playfair, Lambert, and others developed the first graphs that broke away from a geographical representation. In the two centuries since, such charts have become commonplace. With the advent of modern computer tools, creating graphs from data involves trivial effort. In fact, it has probably become too easy. Graphs are often produced without thought for their main purpose: to enlighten and inform the reader.

Tufte defines graphical excellence as "that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."[76] He promotes several rules that will help you avoid the most common pitfalls, which I summarize below:

- "*Above all else show the data":* Make sure the graphs highlight your results instead of hiding them as many graphs do. Use the graphs to tell stories about your analysis, and make those stories leap out at the reader. 18 32

- "*Maximize the data-to-ink ratio, within reason":* Tufte defines the data-ink ratio as the amount of ink that conveys information divided by the total amount of ink used for the graph. Many graphs have low data-ink ratios, and you should avoid this pitfall. To maximize the data-ink ratio, "erase non-data-ink and redundant data-ink."

- "*Eliminate chart junk":* Remove clashing fill patterns (Moiré patterns), which are hard on the eyes. Eliminate extraneous gridlines, which obscure the data. Delete graphical elements that decorate but convey no information, such as 3-dimensional (3-D) chart elements on 2-D graphs.

  This latter mistake is a common one, particularly in "slick" brochures and company annual reports. Although there are legitimate uses for 3-D graphs, a bar chart is not one of them. I adapted **Figure 35.1** from the 1999 annual report of a company with \$6 billion in sales in 1999.

Figure 35.1: The wrong way to make a bar chart

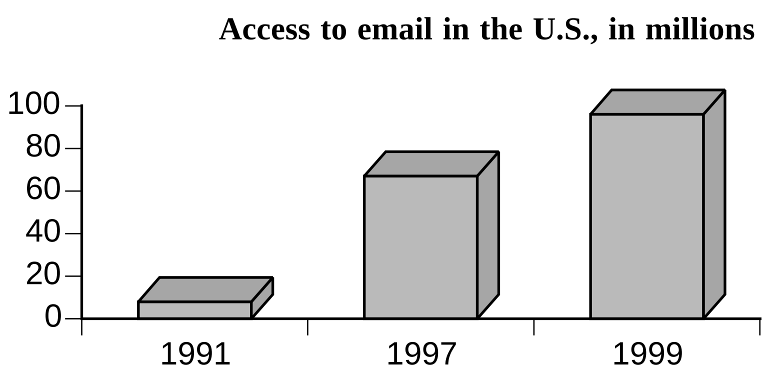
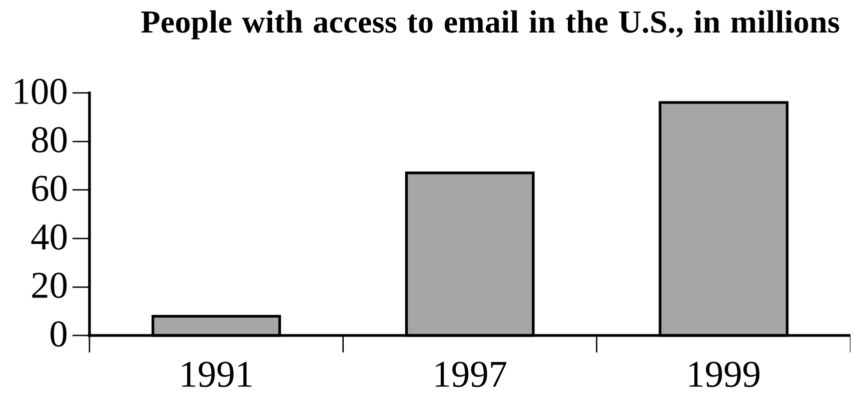


Figure 35.1 shows access to email in the U.S. using 3-D bars. Because of the 3-D perspective, it is impossible to extract the data accurately from the graph. **Figure 35.2** shows how the graph should look for greatest accuracy.

Figure 35.2: A better way to make a bar chart



- "*Maximize data density" (numbers per unit area), within reason:* Typical graphs in newspapers and journals, based on Tufte's review of hundreds of charts published 1970-1980, have data densities of 5 to 20 numbers per square inch (0.8 to 3 numbers per square centimeter). The densest graph Tufte examined (a map of galaxies contributed by some astronomers) had a data density of 110,000 numbers per square inch (17,000 numbers per square centimeter); the least dense graph he considered had a data density of 0.15 numbers per square inch (0.02 numbers per square centimeter). Maps tend to have high data density; badly designed charts tend to have the lowest.

  When things get really bad, a graph may contain only one data point. I encountered such an appalling graph on an airplane flight from San Francisco to Washington, DC on June 8, 1999. It showed the size of the market for some electronic gadget or other, and it looked like **Figure 35.3**:

  Figure 35.3: The market for an electronic gadget

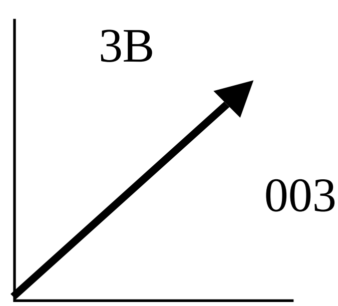
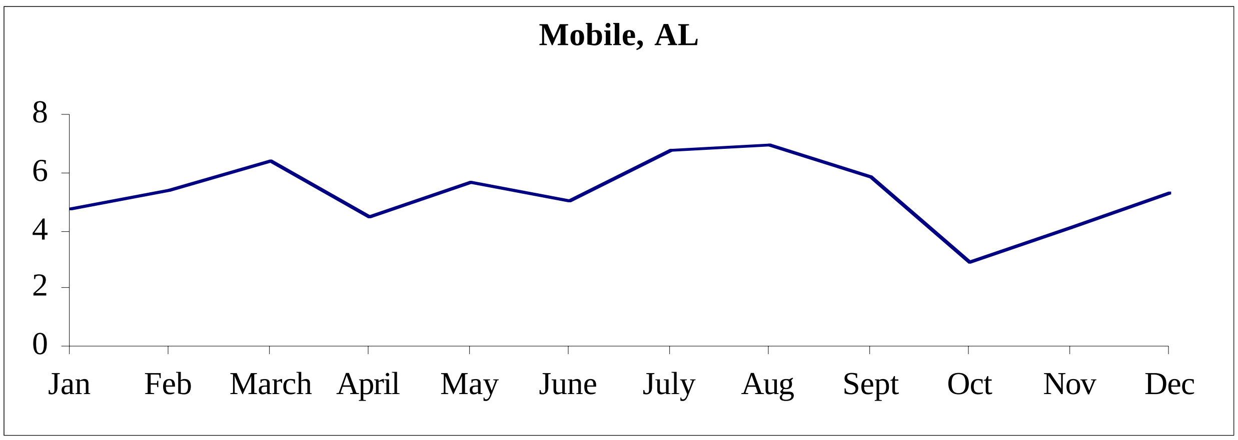


  Figure 35.3 was all that appeared on a 17" diagonal TV screen, which represents a data density of 0.008 numbers per square inch (0.001 numbers per square centimeter), probably an all-time low. It gives no indication of the current size of the market except that the arrow going up and to the right implies that the market is growing from today until 2003. Avoid this blunder and maximize data density.

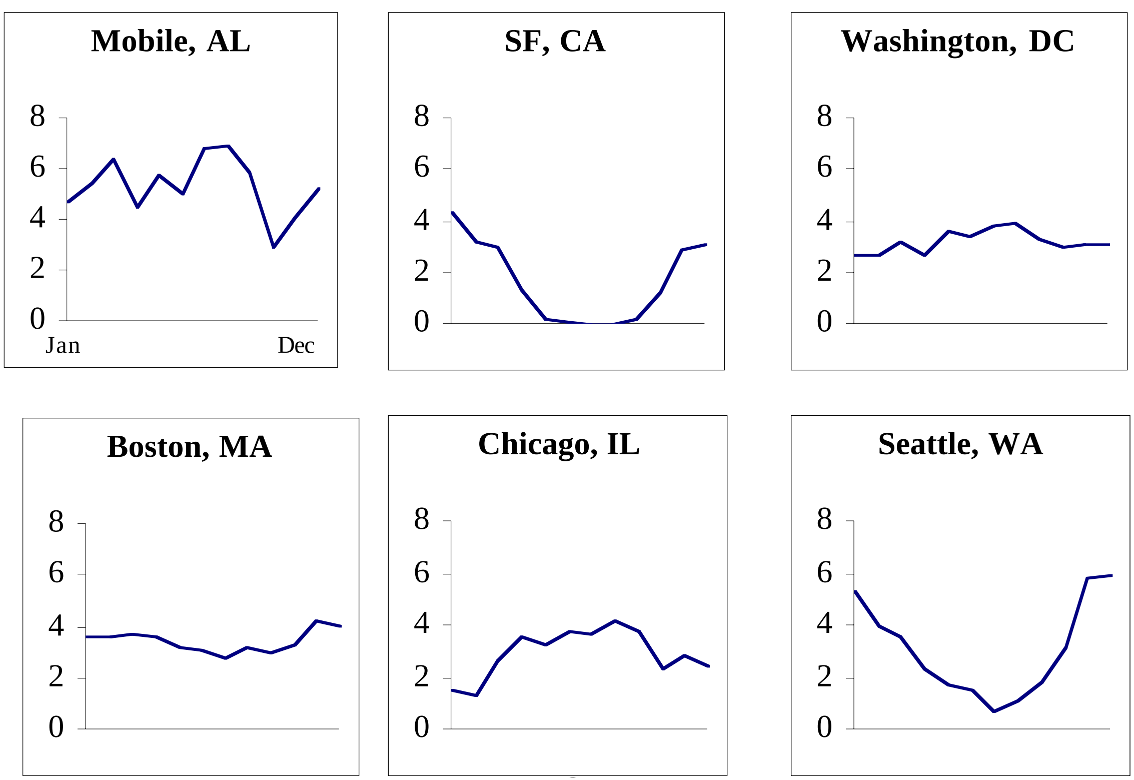
- "*Use small multiples*": One important way to maximize data density is to create what Tufte calls small multiples, defined as "a series of graphics, showing the same combination of variables, indexed by changes in another variable."[77] Let's say you plot monthly rainfall for Mobile, Alabama, as shown in **Figure 35.4**. If you have data on rainfall for many cities that you'd like to compare, you can shrink them and plot them side by side, as shown in **Figure 35.5**. You need not reproduce all the details of the chart labels on every chart. Once the reader has learned the "key" to reading the chart, small multiples like these allow quick and easy comparison of large amounts of data.

Figure 35.4: Average monthly rainfall for Mobile, Alabama, 1961-1990 (inches)



Source: *Statistical Abstract of the U.S. 1997*, Table 397. Normal Monthly and Annual Precipitation–Selected Cities (based on 1961-1990 data).

Figure 35.5: Average monthly rainfall for selected cities, 1961-1990 (inches)



- *"Revise and edit":* All work improves with revisions. Leave enough time to get feedback and make your graphs better.

I add one more rule of my own, which is *Make figures stand alone*. Figures are often copied separately from the report in which they appear, to allow the reader to trace your figure back to the source. Put the date and file name on the figure, at least when it's in draft form. Make sure the final version contains some reference to the report in which it appears or to the author's contact information. Figures are often copied in black and white even if the original graph is in color. Make sure a black and white copy conveys the same message as the original, and give the reader the report and author's contact information on the graph as a backup in case the colors don't come through in the reader's copy.

APPLYING THESE RULES

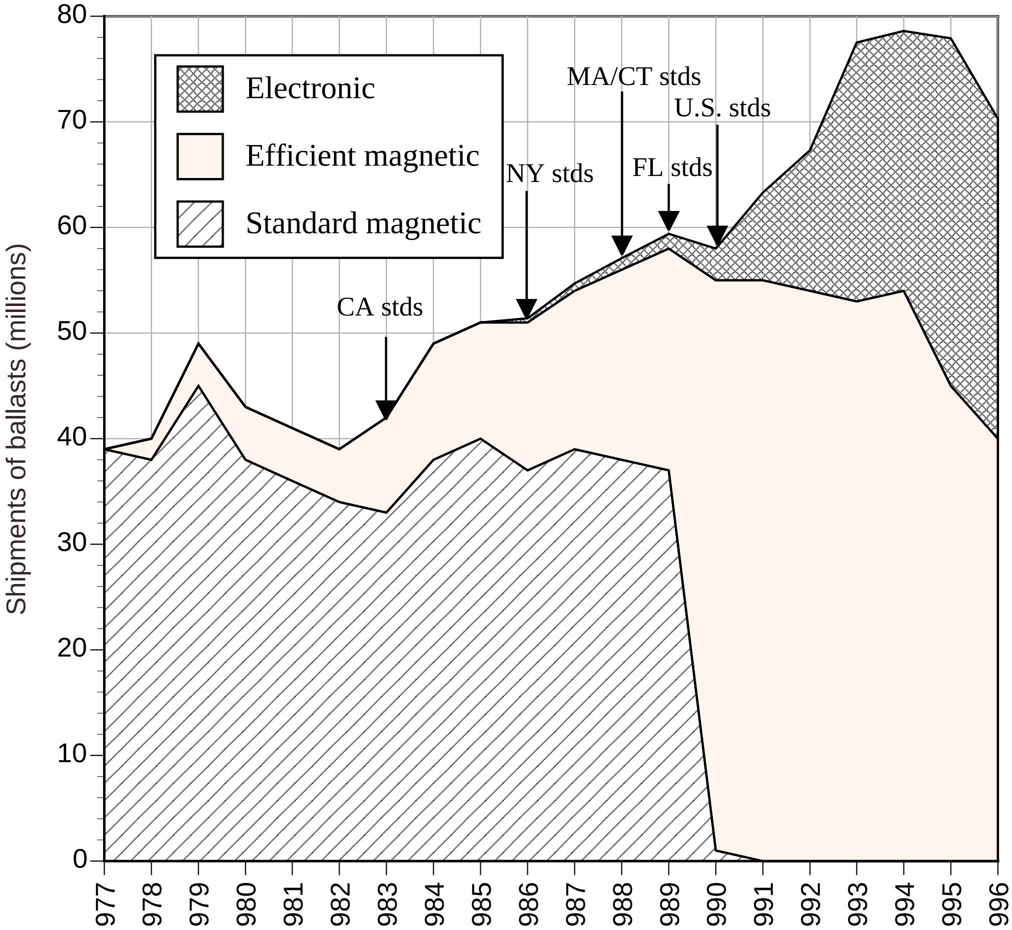
Let's look at a figure I created in the past and make it better using these rules. **Figure 35.6** shows a graph of the historical annual sales of fluorescent lamp ballasts for U.S. commercial buildings that appeared in a 1997 report.[78]

Notice that the patterns used in the area chart seem to "vibrate" when you stare at them. This is what Tufte calls the Moiré effect, and it's best avoided. The gridlines are extraneous, and the legend can easily be eliminated. The years along the X-axis are a bit hard to read because they are turned 90 degrees from horizontal. Some of the text labels are in Times font while the rest of the labels, titles, and text are in Helvetica. The footnote is descriptive and complete, but the figure still lacks the contact information for the author of the report and the file name of the graph.

Figure 35.7 is the improved version of the graph, which has a higher data-to-ink ratio than does Figure 35.6 and is easier to read. I eliminated the Moiré vibration by using shades of gray and white and eliminated the legend by using labels directly on top of the appropriate areas of the graph. I removed the gridlines as well as the box around the graph and reduced the number of tick marks on the Y-axis to only the major ones. I shrank the type size of the years along the X-axis and shifted to a different convention of representing the years ('94 instead of 1994) for all years but the first and last (another choice would have been to lay out the graph in landscape format so that the X-axis was longer). I changed all text to Helvetica font. Finally, I added author's contact information, the file name, and the dates of creation and last modification for the file.

Figure 35.6: A graph that needs work

U.S. shipments of fluorescent ballasts for the U.S. commercial sector



Note: Because they are not typically used in the commercial sector, the following ballast types are omitted from the data presented in this figure: low-power-factor ballasts (shipments of 24.2 million in 1996), 1500 MA high-power-factor ballasts (0.2 million in 1996), and ballasts included in the "All other corrected power factor" category of the Census Bureau data. Ballast imports are not included because they were not disaggregated by ballast type in the census data. Arrows indicate the years in which state standards and, finally, federal standards took effect, preventing the sale of standard magnetic ballasts.

Sources: Koomey et al. (1995); Census Bureau MQ36C(94)-5, Table 3; Census Bureau MQ36C(96)-5, Tables 3 and 4

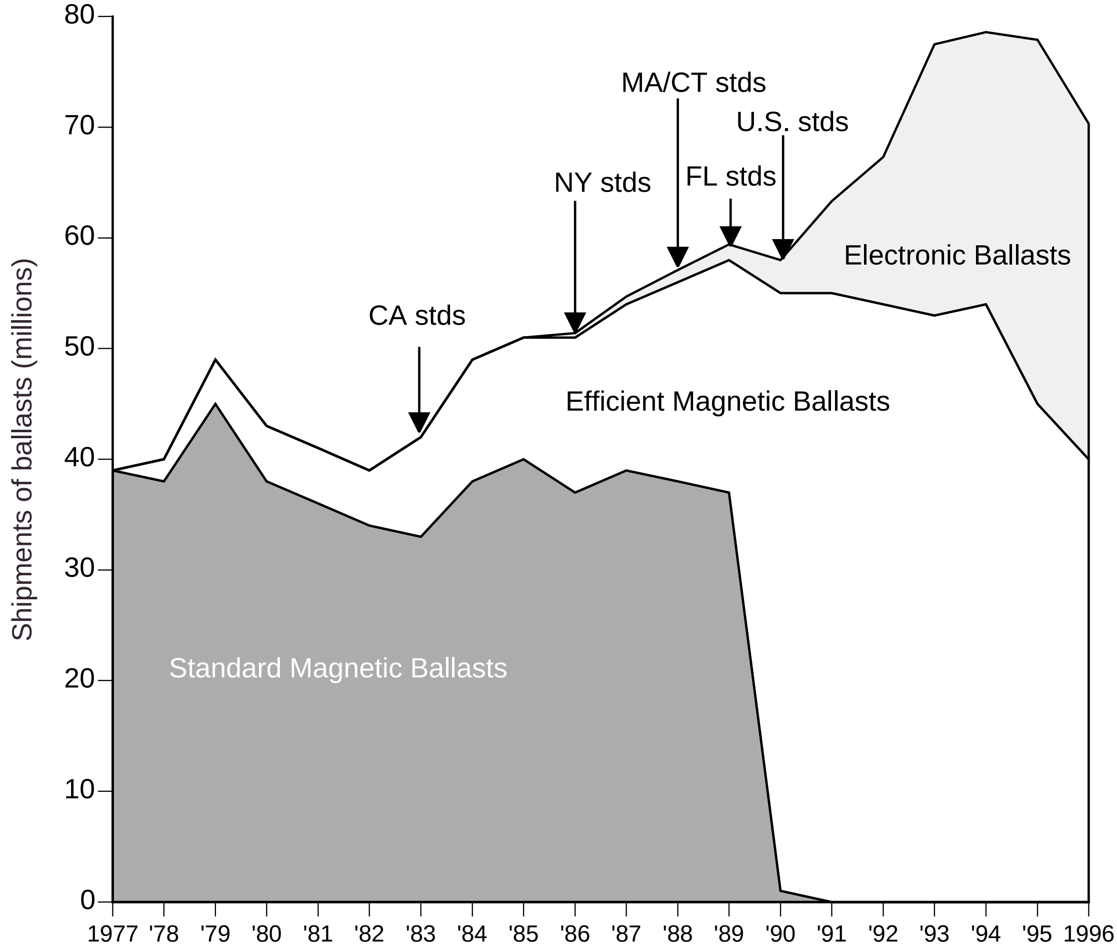
CONCLUSIONS

Tufte reminds us that rules should not be followed slavishly and that our focus should be on the key goal: "to give visual access to the subtle and the difficult."[79] Graphics should serve your arguments. If they don't support your analysis in a direct and understandable way, they should be ruthlessly deleted or drastically modified. As Tufte says, graphs and figures that don't contribute to the "revelation of the complex"[80] have no place in your work.

Read Tufte's books. Review his guidelines periodically, and apply them to your current work. Small changes in graphics design can vastly improve the effectiveness of your figures and graphs.

Figure 35.7: An improved graph

U.S. shipments of fluorescent ballasts for the U.S. commercial sector



Note: Because they are not typically used in the commercial sector, the following ballast types are omitted from the data presented in this figure: low-power-factor ballasts (shipments of 24.2 million in 1996), 1,500-MA high-power-factor ballasts (0.2 million in 1996), and ballasts included in the "All other corrected power factor" category of the Census Bureau data. Ballast imports are not included because they were not disaggregated by ballast type in the census data. Arrows indicate the years in which state standards and, finally, federal standards took effect, preventing the sale of standard magnetic ballasts.

Sources: Koomey et al. (1995); Census Bureau MQ36C(94)-5, Table 3; Census Bureau MQ36C(96)-5, Tables 3 and 4.

Contact: Jonathan Koomey, JGKoomey@lbl.gov. <http://enduse.lbl.gov> Filename: Final Ballast Charts. Date of creation: February 1995. Date of last modification: October 1997.

NOTE TO THE TYPESETTER: PLEASE MAKE SURE THAT THIS IMPROVED VERSION OF THE GRAPH FOLLOWS IMMEDIATELY AFTER THE BAD VERSION, THEN FINISH WITH THE CONCLUSIONS AND THE FOLLOWING QUOTE:

> **Visual representations of evidence should be governed by principles of reasoning about quantitative evidence. For information displays, design reasoning must correspond to scientific reasoning. Clear and precise seeing becomes as one with clear and precise thinking.**
>
> **--Edward Tufte--**

36

CREATE GOOD TABLES

A good table is a work of art while a bad one is worse than useless. I have worked to refine my table-making skills over the years, and I am constantly amazed at the perfectly dreadful tables created by authors who should know better. Tables are the heart of a technical report, and care in their creation is one sign of a clear-thinking analyst. A little care can make your tables a resource that your readers keep as a reference for many years.

I follow several key principles in creating a table:

- *Make them stand alone*: Tables are often copied separately from the report from which they came. Make sure the reader can trace your table back to the source (or at least to the author) and that your calculations are explained clearly in the footnotes.

- *Create extensive footnotes*: This aspect of documentation is crucial. Footnotes to the table should explain the methodology, list the key sources, and describe how each source was used. Any intelligent person should be able to recreate the table from your footnotes. Tufte says that footnotes indicate "care and craft," and readers will be grateful when you create them. 33

  Of course, for tables appearing in more popular publications (like newspapers) the standards for footnote documentation are often lower, but even if the footnotes don't appear in the printed version, the person who created the table should still have the footnotes filed away in case there's a need to check the calculations later.

- *"Maximize the data-ink ratio," within reason*: Edward Tufte's guidance is just as relevant for tables as for graphs. Make sure that every part of the table conveys information. Avoid extraneous gridlines (only include them when they enhance reader comprehension). Use carefully chosen white space to separate and group results. 35

- *Some redundancy is good*: Unlike in graphs, where redundancy should generally be avoided, showing additional columns or rows for totals and subtotals is essential. These ostensibly redundant sums help the reader follow the calculation and assure that the table is internally consistent. The sums can also help in troubleshooting your own calculations.

  Another important redundancy relates to percentages. If you have a column of percentages, label the column heading "% of total" AND put a percent symbol in every cell (e.g., 1.37%, not 1.37). Sometimes people make mistakes in their calculations and forget to divide by 100 to make a real percentage. By enforcing this redundancy, you guarantee that readers will interpret your results correctly and you'll save them time. Having the % symbol in every cell makes the meaning unambiguous.

  Finally, all numbers less than zero should have a leading zero (e.g., 0.35, NOT .35). Copies become smudged and decimal points can be obscured. With a leading zero, there will never be ambiguity about the size of the number in question.

- *Enhance readability with font, style, and orientation choices*: In her *Non-Designer's Design Book*, Robin Williams demonstrates the principles of proximity, alignment, repetition, and contrast, and applies them to everyday design problems.

Use spacing, justification (center, left, or right), grid lines, italics, and bold-facing to implement these principles within the table. Align decimal points within each column to add even more information.

Choose a font that is easy on the eyes and doesn't deteriorate as a document is copied. Tufte recommends using the font used in telephone books (Bell Centennial) whenever you have tables that have high information density because the phone companies have researched this topic extensively, and there's no point in reinventing the wheel.

- *Enhance reader comprehension*: The single most important tool to increase usefulness of tabular data is to add rows or columns containing *indices* relative to a 18 27
total or subtotal, which will show the relative importance of specific numbers relative to these totals (see the examples below). Most readers will calculate such indices themselves in their effort to understand the table. Do it for them, and they will thank you for it.

- *Normalize data to facilitate comparisons:* It's often helpful to normalize results on 27
a per capita or per GDP basis, so the reader can compare them to more familiar quantities.

- *Create the table first; then, write your text from the tables*: Creating the tables helps 34
you think through the analysis and ensures that your logic is both correct and reproducible. Ask yourself questions about the analysis and check your work. If you can't explain it in writing, then you don't understand it. Clear tables and clear writing come from clear thinking.

- *Avoid spurious precision:* Only show the appropriate number of significant figures. Adding extra decimal places is distracting and pernicious. For example, results of calculations that use inputs reliable to one or two significant figures (e.g., 100 divided by 51) should not then be reported to five decimal places (1.93453). The correct value to use in this case is 2. The calculation is only as accurate as the least accurate input parameter (for more details on treating significant figures, see <http://www.chem.sc.edu/faculty/morgan/sigfigs>.

- *Revise and edit*: Show your tables to colleagues before publication, and you will surely learn something. What is clear and obvious to you may not be so obvious to them. Give yourself time to revise and improve the tables.

A few specific examples will help illustrate some of these principles (A Microsoft Excel file [improvingtables.xls] containing the original and improved tables can be downloaded at <http://www.numbersintoknowledge.com>. The first is taken from a book written by a physics professor with many publications to his credit, and the publisher was one of the largest in the U.S.[81] **Table 36.1** shows how the original table appeared.

The first thing to notice is that the footnote is ambiguous. Is the total paper component split into 10 to 15% newspapers and 40 to 35% packaging, or is total paper split into 5 to 7.5% newspapers (50% times 10% to 15%) and 45 to 42.5% packaging? The former interpretation is more likely, so that's what I assumed for purposes of this example, but it is impossible to know for sure.

Table 36.1: The table as it appeared in the book

Contents of American Landfills

MATERIAL	PERCENT OF VOLUME OCCUPIED
Paper	50*
Plastic	10
Organic (includes yard waste)	13
Metal	6
Glass	1
Miscellaneous (tires, rubber, debris)	20

*10 to 15 percent of the total volume is newspapers; the rest is packaging.

Table 36.2 shows what the improved table looks like. To make it, I first added a "Total" row, to remind readers that the total adds to 100%. I added gridlines to separate the main sections of the table, italicized the column headings, and changed the column headings to have lower case letters after the first letter of each word. I added the % symbol after each number to eliminate any ambiguity (this change was less important here than it would be if some of the values were less than 1). I incorporated the information contained in the original footnote into the table itself by showing subtotals under the "paper" category for newspaper and packaging. I right-justified these subtotals and reduced the type size to 9 points, to make it clear that these numbers are subordinate to the paper category and different than the other numbers in the table. I added a footnote giving the name of the source for the information and moved some of the qualifications for two of the waste categories into two other footnotes. This last change makes the additional information accessible to the interested reader, but removes it from the main table where it distracts the general reader from quickly grasping the main results.

Table 36.2: An improved table

Contents of American Landfills[1]

Material	*Percent of Landfill Volume Occupied*
Total paper	50%
Newspaper	10 to 15%
Paper packaging	40 to 35%
Plastic	10%
Organic[2]	13%
Metal	6%
Glass	1%
Miscellaneous[3]	20%
Total	100%

[1]Source: William Rathje, University of Arizona.

[2]Organic waste includes yard waste.

[3]Miscellaneous includes tires, rubber, and debris.

The second example, shown in **Table 36.3** is taken from one of my group's technical reports.[82] This table slipped through our quality control, but I'm making amends by dissecting it for this example. Note the extraneous gridlines and the lack of indices that would help the reader focus on the key results from the table.

Table 36.4 shows how to revise the table. I eliminated all but the essential gridlines and included indices relative to 1983 (which show how much shipments and imports grew from 1983 to 1994). I also added a column with imports normalized as the percentage of U.S. shipments, to help the reader understand how this important statistic changed over time. In addition, I included an additional row with the annual average percentage change from 1983 to 1994 because this summary statistic is helpful for understanding long-term trends in the data. Finally, I added information to the footnotes to identify the contact person, the file name, and the dates of creation and last modification for the file.

Table 36.5 shows a third example, taken from another technical report that my group worked on.[83] In this case, the client insisted on a particular format for the table, but **Table 36.6** shows how I'd change the table if it were up to me (it's not bad, but it could be better). To improve it, I added a column containing the average annual percentage change from 1973 through 1997. I also added a row with an index relative to 1973, which (for example) shows at a glance that primary energy use in the U.S. grew 24% over this period. Finally, I added contact information, so readers with questions know whom to approach. The more detailed information with the file name and date of creation and last modification makes sense for a technical table but would be excessive for a summary table.

Table 36.5 appeared in a summary chapter of a larger report, so the recommended revisions are slightly different than if the table were to appear in a more detailed technical chapter (**Table 36.7** shows a version that would be appropriate in that situation). The same information on primary energy use in the different sectors is shown in that table, supplemented by several indices. First, the table shows energy use by sector expressed as a fraction of the total. Next, it shows indices relative to 1973, making growth over the analysis period readily apparent. The table also includes average annual percentage rates of change since 1973 for the sectors and the total. Finally, the additional detail about the file name and dates of creation and last modification are appropriate here.

Although some aspects of creating good tables are matters of style and preference, the principles described above should remain foremost in your mind, because they are valid for almost every table. Readers should treasure your tables for their accuracy and completeness. Sloppy tables sap your credibility—avoid them at all costs.

> **There are two ways of constructing a software design [or a table]: one way is to make it so simple that there are obviously no deficiencies; the other is to make it so complicated that there are no obvious deficiencies.**
>
> **--C. A. R. Hoare--**

Table 36.3: A table in one of my group's reports that needs revision

U.S. Shipments and Imports of Lamps, 1983-1994

Year	U.S. Shipments (millions of lamps)	Imports (millions of lamps)
1983	3615.9	560.9
1984	3723.4	748.7
1985	3472	862.7
1986	3421.3	920.6
1987	3399.4	999.8
1988	3510.2	1130.8
1989	3429.5	1024
1990	3318.5	1051
1991	3297.5	Data unavailable from Census Bureau
1992	3422.1	Data unavailable from Census Bureau
1993	3564.3	1372.6
1994	3563.3	1577.8

Note: "U.S. shipments" refers to total shipments by manufacturers located within the U.S., including units to be exported. Cold-cathode fluorescent lamps are excluded from the U.S. shipment data; Christmas tree lights are excluded from U.S. shipments as well as imports.

Source: Census Bureau current industrial reports MQ36B, various years.

Table 36.4: Table 36.3, improved and revised

U.S. Shipments and Imports of Lamps, 1983-1994

Year	*U.S. Shipments* Millions of lamps	*U.S. Shipments* Index 1983 = 1.00	*Imports* Millions of lamps	*Imports* Index 1983 = 1.00	*Imports as a percentage of U.S. shipments*
1983	3,616	*1.00*	561	*1.00*	16%
1984	3,723	*1.03*	749	*1.33*	20%
1985	3,472	*0.96*	863	*1.54*	25%
1986	3,421	*0.95*	921	*1.64*	27%
1987	3,399	*0.94*	1,000	*1.78*	29%
1988	3,510	*0.97*	1,131	*2.02*	32%
1989	3,430	*0.95*	1,024	*1.83*	30%
1990	3,319	*0.92*	1,051	*1.87*	32%
1991	3,298	*0.91*	NA	*NA*	NA
1992	3,422	*0.95*	NA	*NA*	NA
1993	3,564	*0.99*	1,373	*2.45*	39%
1994	3,563	*0.99*	1,578	*2.81*	44%
Average annual percentage change 1983 through 1994	-0.1%		9.9%		10.0%

Note: "U.S. shipments" refers to total shipments by manufacturers located within the U.S., including units to be exported. Cold-cathode fluorescent lamps are excluded from the U.S. shipment data; Christmas tree lights are excluded from U.S. shipments as well as imports.

NA = Not Available

Source: U.S. Census Bureau current industrial reports MQ36B, various years

Contact: Jonathan Koomey, JGKoomey@lbl.gov. http://enduse.lbl.gov/

Filename: shipmentsandimports.xls

Date of creation: February 1996

Date of last modification: November 1997

Table 36.5: A table that needs revision from another of our technical reports

Primary energy use in quads: 1973–1997

	1973	1986	1990	1995	1997
Buildings	24.1	26.9	29.4	32.1	33.7
Industry	31.5	26.6	32.1	34.5	32.6
Transportation	18.6	20.8	22.6	24.1	25.5
Total	74.2	74.3	84.1	90.7	91.8

Source: Energy use estimates for 1973-95 come from EIA (1996a, Table 1.1, p. 39). Energy use estimates for 1997 come from EIA (1996c).

Table 36.6: Table 36.5 improved for use in a summary report

Primary energy use in quads: 1973–1997

	1973	*1986*	*1990*	*1995*	*1997*	*Average annual change 1973-1997 % per year*
Buildings	24.1	26.9	29.4	32.1	33.7	1.4%
Industry	31.5	26.6	32.1	34.5	32.6	0.1%
Transportation	18.6	20.8	22.6	24.1	25.5	1.3%
Total	74.2	74.3	84.1	90.7	91.8	0.9%
Index (1973 = 1.00)	*1.00*	*1.00*	*1.13*	*1.22*	*1.24*	

Source: Energy use estimates for 1973-95 come from EIA (1996a, Table 1.1, p. 39). Energy use estimates for 1997 come from EIA (1996c).

Contact: Jonathan Koomey, JGKoomey@lbl.gov. http://enduse.lbl.gov/

Table 36.7: Table 36.5 improved for use in a detailed technical report

Primary energy use: 1973–1997

	1973	1986	1990	1995	1997
	Quadrillion Btus of primary energy				
Buildings	24.1	26.9	29.4	32.1	33.7
Industry	31.5	26.6	32.1	34.5	32.6
Transportation	18.6	20.8	22.6	24.1	25.5
Total	74.2	74.3	84.1	90.7	91.8
	Sectoral energy as a percent of total				
Buildings	32%	36%	35%	35%	37%
Industry	42%	36%	38%	38%	36%
Transportation	25%	28%	27%	27%	28%
Total	100%	100%	100%	100%	100%
	Index 1973 = 1.00				
Buildings	1.00	1.12	1.22	1.33	1.40
Industry	1.00	0.84	1.02	1.10	1.03
Transportation	1.00	1.12	1.22	1.30	1.37
Total	1.00	1.00	1.13	1.22	1.24
	Average annual percentage change since 1973				
Buildings	0%	0.8%	1.2%	1.3%	1.4%
Industry	0%	-1.3%	0.1%	0.4%	0.1%
Transportation	0%	0.9%	1.2%	1.2%	1.3%
Total	0%	0.0%	0.7%	0.9%	0.9%

Source: Energy use estimates for 1973-95 come from EIA (1996a, Table 1.1, p. 39). Energy use estimates for 1997 come from EIA (1996c).

Contact: Jonathan Koomey, JGKoomey@lbl.gov. http://enduse.lbl.gov/

Filename: primaryenergyuse.xls

Date of creation: January 1999

Date of last modification: December 1999